

Amendment to the Claims

This listing of claims will replace all prior versions, and listings, of claims in the application:

1 Claims 1-18 (canceled).

1 19. (currently amended) A system ~~according to Claim 18, further for~~
2 providing efficient document scoring of concepts within and clustering of
3 documents in an electronically-stored document set, comprising:
4 [[the]] a database electronically storing a document set;
5 a scoring module ~~evaluating the score~~ scoring a document in the
6 electronically-stored document set, comprising:
7 a frequency submodule determining a frequency of occurrence of
8 at least one concept within a document;
9 a concept weight submodule analyzing a concept weight reflecting
10 a specificity of meaning for the at least one concept within the document, wherein
11 the concept weight is based on a number of terms for the at least one concept;
12 a structural weight submodule analyzing a structural weight
13 reflecting a degree of significance based on structural location within the
14 document for the at least one concept;
15 a corpus weight submodule analyzing a corpus weight inversely
16 weighing a reference count of occurrences for the at least one concept within the
17 document;
18 a scoring evaluation submodule evaluating a score to be associated
19 with the at least one concept as a function of a summation of the frequency,
20 concept weight, structural weight, and corpus weight in accordance with the
21 formula:

22
$$S_i = \sum_{j \rightarrow n}^j f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

23 where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises
24 the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$
25 comprises the corpus weight for occurrence j of concept i ;
26 a vector submodule forming the score assigned to the at least one
27 concept as a normalized score vector for each such document in the
28 electronically-stored document set; and
29 a determination submodule determining a similarity between the
30 normalized score vector for each such document as an inner product of each
31 normalized score vector;
32 a clustering module grouping the documents by the score into a plurality
33 of clusters, comprising:
34 a selection submodule selecting a set of candidate seed documents
35 from the electronically-stored document set;
36 a cluster seed submodule identifying seed documents by applying
37 the similarity to each such candidate seed document and selecting those candidate
38 seed documents that are sufficiently unique from other candidate seed documents
39 as the seed documents;
40 an identification submodule identifying a plurality of non-seed
41 documents;
42 a comparison submodule determining the similarity between each
43 non-seed document and a cluster center of each cluster; and
44 a clustering submodule assigning each such non-seed document to
45 the cluster with a best fit, subject to a minimum fit;
46 a threshold module relocating outlier documents, comprising determining
47 the similarity between each of the documents grouped into each cluster based on
48 the center of the cluster and the scores assigned to each of the at least one
49 concepts in that document, dynamically determining a threshold for each cluster
50 as a function of the similarity between each of the documents, and identifying and
51 reassigning each of the documents with the similarity falling outside the
52 threshold; and

53 a processor to execute the modules and submodules.

1 20. (previously presented) A system according to Claim 19, further
2 comprising:
3 the concept weight module evaluating the concept weight in accordance
4 with the formula:

$$5 \quad cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij})^{\frac{1}{3}}, & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}])^{\frac{1}{3}}, & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

6 where cw_{ij} comprises the concept weight and t_{ij} comprises the number of terms for
7 occurrence j of each such concept i .

1 21. (previously presented) A system according to Claim 19, further
2 comprising:
3 the structural weight module evaluating the structural weight in
4 accordance with the formula:

$$5 \quad sw_{ij} = \begin{cases} 1.0, & \text{if } (j \approx \text{SUBJECT}) \\ 0.8, & \text{if } (j \approx \text{HEADING}) \\ 0.7, & \text{if } (j \approx \text{SUMMARY}) \\ 0.5 & \text{if } (j \approx \text{BODY}) \\ 0.1 & \text{if } (j \approx \text{SIGNATURE}) \end{cases}$$

6 where sw_{ij} comprises the structural weight for occurrence j of each such concept i .

1 22. (previously presented) A system according to Claim 19, further
2 comprising:
3 the corpus weight module evaluating the corpus weight in accordance with
4 the formula:

$$5 \quad rw_{ij} = \begin{cases} \left(\frac{T - r_{ij}}{T} \right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

6 where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for
7 occurrence j of each such concept i , T comprises a total number of reference
8 counts of documents in the document set, and M comprises a maximum reference
9 count of documents in the document set.

1 23. (previously presented) A system according to Claim 19, further
2 comprising:

3 a compression module compressing the score in accordance with the
4 formula:

5
$$S'_i = \log(S_i + 1)$$

6 where S'_i comprises the compressed score for each such concept i .

1 24. (currently amended) A system according to ~~Claim 18~~ Claim 19,
2 further comprising:

3 a global stop concept vector cache maintaining concepts and terms; and
4 a filtering module filtering selection of the at least one concept based on
5 the concepts and terms maintained in the global stop concept vector cache.

1 25. (currently amended) A system according to ~~Claim 18~~ Claim 19,
2 further comprising:

3 a parsing module identifying terms within at least one document in the
4 document set, and combining the identified terms into one or more of the
5 concepts.

1 26. (original) A system according to Claim 25, further comprising:
2 the parsing module structuring each such identified term in the one or
3 more concepts into canonical concepts comprising at least one of word root,
4 character case, and word ordering.

1 27. (original) A system according to Claim 25, wherein at least one of
2 nouns, proper nouns and adjectives are included as terms.

Claims 28-30 (canceled).

31. (currently amended) A system according to ~~Claim 18~~ Claim 19,
further comprising:
the similarity submodule calculating the similarity in accordance with the
formula:

$$\cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A| |\vec{S}_B|}$$

where $\cos \sigma_{AB}$ comprises a similarity between a document A and a document B ,
 \vec{S}_A comprises a score vector for document A , and \vec{S}_B comprises a score vector for
document B .

Claims 32-35 (canceled).

36. (currently amended) A computer-implemented method ~~according~~
~~to Claim 35, further for providing efficient document scoring of concepts within~~
~~and clustering of documents in an electronically-stored document set, comprising:~~
~~evaluating the score~~ scoring a document in an electronically-stored
document set, comprising:
determining a frequency of occurrence of at least one concept
within a document;
analyzing a concept weight reflecting a specificity of meaning for
the at least one concept within the document, wherein the concept weight is based
on a number of terms for the at least one concept;
analyzing a structural weight reflecting a degree of significance
based on structural location within the document for the at least one concept;
analyzing a corpus weight inversely weighing a reference count of
occurrences for the at least one concept within the document; and

15 evaluating a score to be associated with the at least one concept as
16 a function of a summation of the frequency, concept weight, structural weight,
17 and corpus weight and in accordance with the formula:

18
$$S_i = \sum_{j=1}^j f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

19 where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises
20 the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$
21 comprises the corpus weight for occurrence j of concept i ;

22 forming the score assigned to the at least one concept as a normalized
23 score vector for each such document in the electronically-stored document set;

24 determining a similarity between the normalized score vector for each
25 such document as an inner product of each normalized score vector;

26 grouping the documents by the score into a plurality of clusters,
27 comprising:

28 selecting a set of candidate seed documents from the
29 electronically-stored document set;

30 identifying seed documents by applying the similarity to each such
31 candidate seed document and selecting those candidate seed documents that are
32 sufficiently unique from other candidate seed documents as the seed documents;

33 identifying a plurality of non-seed documents;

34 determining the similarity between each non-seed document and a
35 center of each cluster; and

36 assigning each non-seed document to the cluster with a best fit,
37 subject to a minimum fit; and

38 relocating outlier documents, comprising:

39 determining the similarity between each of the documents grouped
40 into each cluster based on the center of the cluster and the scores assigned to each
41 of the at least one concepts in that document;

42 dynamically determining a threshold for each cluster as a function
43 of the similarity between each of the documents; and

44 identifying and reassigning each of the documents with the
45 similarity falling outside the threshold.

1 37. (currently amended) A computer-implemented method according
2 to Claim 36, further comprising:
3 evaluating the concept weight in accordance with the formula:

$$4 \quad cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

5 where cw_{ij} comprises the concept weight and t_{ij} comprises the number of terms for
6 occurrence j of each such concept i .

1 38. (currently amended) A computer-implemented method according
2 to Claim 36, further comprising:
3 evaluating the structural weight in accordance with the formula:

$$4 \quad sw_{ij} = \begin{cases} 1.0, & \text{if } (j \approx \text{SUBJECT}) \\ 0.8, & \text{if } (j \approx \text{HEADING}) \\ 0.7, & \text{if } (j \approx \text{SUMMARY}) \\ 0.5 & \text{if } (j \approx \text{BODY}) \\ 0.1 & \text{if } (j \approx \text{SIGNATURE}) \end{cases}$$

5 where sw_{ij} comprises the structural weight for occurrence j of each such concept i .

1 39. (currently amended) A computer-implemented method according
2 to Claim 36, further comprising:
3 evaluating the corpus weight in accordance with the formula:

$$4 \quad rw_{ij} = \begin{cases} \left(\frac{T - r_{ij}}{T} \right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

5 where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for
6 occurrence j of each such concept i , T comprises a total number of reference

7 counts of documents in the document set, and M comprises a maximum reference
8 count of documents in the document set.

1 40. (currently amended) A computer-implemented method according
2 to Claim 36, further comprising:
3 compressing the score in accordance with the formula:
4 $S'_i = \log(S_i + 1)$
5 where S'_i comprises the compressed score for each such concept i .

1 41. (currently amended) A computer-implemented method according
2 to ~~Claim 35~~ Claim 36, further comprising:
3 maintaining concepts and terms in a global stop concept vector cache; and
4 filtering selection of the at least one concept based on the concepts and
5 terms maintained in the global stop concept vector cache.

1 42. (currently amended) A computer-implemented method according
2 to ~~Claim 35~~ Claim 36, further comprising:
3 identifying terms within at least one document in the document set; and
4 combining the identified terms into one or more of the concepts.

1 43. (currently amended) A computer-implemented method according
2 to Claim 42, further comprising:
3 structuring each such identified term in the one or more concepts into
4 canonical concepts comprising at least one of word root, character case, and word
5 ordering.

1 44. (currently amended) A computer-implemented method according
2 to Claim 42, further comprising:
3 including as terms at least one of nouns, proper nouns and adjectives.

1 Claims 45-47 (canceled).

48. (currently amended) A computer-implemented method according
to ~~Claim 35~~ Claim 36, further comprising:

calculating the similarity in accordance with the formula:

$$\cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A| |\vec{S}_B|}$$

where $\cos \sigma_{AB}$ comprises a similarity between a document A and a document B ,

\vec{S}_A comprises a score vector for document A , and \vec{S}_B comprises a score vector for
document B .

Claims 49-51 (canceled).

52. (currently amended) A computer-readable storage medium holding
code for providing efficient document scoring of concepts within and clustering
of documents in an electronically-stored document set, comprising:

code for scoring a document in an electronically-stored document set,
comprising:

code for determining a frequency of occurrence of at least one
concept within a document;

code for analyzing a concept weight reflecting a specificity of
meaning for the at least one concept within the document, wherein the concept
weight is based on a number of terms for the at least one concept;

code for analyzing a structural weight reflecting a degree of
significance based on structural location within the document for the at least one
concept;

code for analyzing a corpus weight inversely weighing a reference
count of occurrences for the at least one concept within the document; and

code for evaluating a score to be associated with the at least one
concept as a function of a summation of the frequency, concept weight, structural
weight, and corpus weight in accordance with the formula:

19
$$S_i = \sum_{j=1}^J f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

20 where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises
21 the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$
22 comprises the corpus weight for occurrence j of concept i ;

23 code for forming the score assigned to the at least one concept as a
24 normalized score vector for each such document in the electronically-stored
25 document set;

26 code for determining a similarity between the normalized score vector for
27 each such document as an inner product of each normalized score vector;

28 code for grouping the documents by the score into a plurality of clusters,
29 comprising:

30 code for selecting a set of candidate seed documents from the
31 electronically-stored document set;

32 code for identifying seed documents by applying the similarity to
33 each such candidate seed document and selecting those candidate seed documents
34 that are sufficiently unique from other candidate seed documents as the seed
35 documents;

36 code for identifying a plurality of non-seed documents;

37 code for determining the similarity between each non-seed
38 document and a center of each cluster; and

39 code for assigning each non-seed document to the cluster with a
40 best fit, subject to a minimum fit; and

41 code for relocating outlier documents, comprising:

42 code for determining the similarity between each of the documents
43 grouped into each cluster based on the center of the cluster and the scores
44 assigned to each of the at least one concepts in that document;

45 code for dynamically determining a threshold for each cluster as a
46 function of the similarity between each of the documents; and

code for identifying and reassigning each of the documents with
the similarity falling outside the threshold.

53. (currently amended) An apparatus for providing efficient
document scoring of concepts within and clustering of documents in an
electronically-stored document set, comprising:

means for scoring a document in an electronically-stored document set,
comprising:

means for determining a frequency of occurrence of at least one
concept within a document;

means for analyzing a concept weight reflecting a specificity of
meaning for the at least one concept within the document, wherein the concept
weight is based on a number of terms for the at least one concept;

means for analyzing a structural weight reflecting a degree of
significance based on structural location within the document for the at least one
concept;

means for analyzing a corpus weight inversely weighing a
reference count of occurrences for the at least one concept within the document;
and

means for evaluating a score to be associated with the at least one
concept as a function of a summation of the frequency, concept weight, structural
weight, and corpus weight in accordance with the formula:

$$S_i = \sum_{j \rightarrow n} f_j \times cw_j \times sw_j \times rw_j$$

where S_i comprises the score, f_j comprises the frequency, $0 < cw_j \leq 1$ comprises
the concept weight, $0 < sw_j \leq 1$ comprises the structural weight, and $0 < rw_j \leq 1$
comprises the corpus weight for occurrence j of concept i ;

means for forming the score assigned to the at least one concept as a
normalized score vector for each such document in the electronically-stored
document set;

27 means for determining a similarity between the normalized score vector
28 for each such document as an inner product of each normalized score vector;
29 means for grouping the documents by the score into a plurality of clusters,
30 comprising:
31 means for selecting a set of candidate seed documents from the
32 electronically-stored document set;
33 means for identifying seed documents by applying the similarity to
34 each such candidate seed document and selecting those candidate seed documents
35 that are sufficiently unique from other candidate seed documents as the seed
36 documents;
37 means for identifying a plurality of non-seed documents;
38 means for determining the similarity between each non-seed
39 document and a center of each cluster; and
40 means for assigning each non-seed document to the cluster with a
41 best fit, subject to a minimum fit; and
42 means for relocating outlier documents, comprising:
43 means for determining the similarity between each of the
44 documents grouped into each cluster based on the center of the cluster and the
45 scores assigned to each of the at least one concepts in that document;
46 means for dynamically determining a threshold for each cluster as
47 a function of the similarity between each of the documents; and
48 means for identifying and reassigning each of the documents with
49 the similarity falling outside the threshold.